Issues in corpus development for multi-party multi-modal task-oriented dialogue

Susan Robinson, Bilyana Martinovski, Saurabh Garg[†], Jens Stephan, and David Traum

USC Institute for Creative Technologies
13274 Fiji Way
Marina del Rey, CA 90292, USA
{robinson, martinovski, stephan, traum}@ict.usc.edu

Abstract

This paper describes the development of a multi-modal corpus based on multi-party multi-task driven common goal oriented spoken language interaction. The data consists of approximately 10 hours of audio human simulation radio data and nearly 5 hours of video and audio face-to-face sessions between human trainees and virtual agents.

1. Introduction

Corpus-based approaches to dialogue have become an increasingly important part of dialogue agent design, providing a scope of the real issues that need to be dealt with in order to engage in natural dialogue with humans, as well as providing the basic data for statistical methods for language processing. While there has been much work on dyadic conversation, both casual (e.g., Switchboard (Jurafsky et al., 1998)) and task oriented (e.g., Maptask (Carletta et al., 1996) and Trains (Heeman and Allen, 1994)), there has been much less emphasis on multiparty dialogue. The mission rehearsal exercise project (MRE) (Swartout et al., 2004) has the ambitious aim of virtual reality training of a decision-maker in a multi-party mission-related setting. This activity differs from those found in existing available corpora in several aspects:

- there are multiple participants communicating in a variety of modalities (face to face, over the radio)
- talk is centered around a central "mission" and actions taken to further that mission and cope with issues that arise while carrying out that mission. Talk is not a single short dialogue, but a series of "dialogue episodes", each of which may be between different participants, using different modalities, and separated by periods of non-talk, while acting in the environment or waiting for results of previous interaction.
- The participants involved have a rich social structure, including hierarchical status and differing areas of competence and responsibility.
- There are domain and genre-specific aspects of interaction, including special terms and phrases.

These differences not only necessitate developing corpora as a resource for analysis in such domains, they require some special features in corpus design and annotation. This paper describes our current status in developing such a corpus. In the next section we describe the domain of the MRE system and describe features of the data types comprising our corpus. In section 3., we describe preliminary observations on the corpus, and in section 4., we discuss future work.

2. MREC Domain Description

The specific activities we are concerned with are military mission-related training activities. The MRE virtual reality training project is set within the specific domain of decision-making for a platoon-leader in a peace-keeping mission in Bosnia. The system is set in a room with 150 degree field of view screen with 3-D immersive sound. The trainee plays the role of a lieutenant who is given the mission to assist another platoon in a weapons inspection. Enroute, he encounters an accident between an army vehicle and civilian car, with an injured boy lying wounded on the ground. He must decide whether to help the boy, continue on his assigned mission, or split his forces and attempt both actions. He may interact with a number of visually accessible characters (the sergeant, a medic, the boy's mother), as well as characters over the radio, such as the base and the increasingly desperate commander of the other platoon.

2.1. MRE Bosnia

In order to get a good sense of how a trainee might react within a near-term realistic situation, it was important to constrain the range of responses to something approximating the anticipated experience. While initially this data was gathered with a wizard of oz interface, we have increasingly relied on data from actual runs of the MRE system itself to further develop both task inclusion and improvement of the dialogue system. The average length of these dialogues is 618 total words, broken into 128 utterances, on average (49 human utterances and 79 agent utterances). Table 1 shows the average amounts of MRE dialogue per recording session (each consisting of several dialogues with different participants). Although the human lieutenant talks more than any one character, it is slightly unusual that the agents together talk more than the Lieutenant. One can also notice a trend for the humans to talk less and the agents to speak for a higher percentage of the total interaction.

The multi-participant, multimodal nature of the MRE scenario can yield complex overlapping dialogues, as in fig-

[†]National Center for Ecological Analysis and Synthesis, Suite 300, 735 State Street, Santa Barbara, CA 93101-3351,sgarg@ecoinformatics.org

date	human	agent	combined
Mar	410	350	760
May	260	449	709
Aug	198	314	512
Nov 6	153	358	511
Nov 20	257	525	782
Dec	192	349	541

Table 1: Average dialogue size by collection period

ure 1, a segment from dialogue between a human trainee (Lt) and virtual agents during one of the MRE test runs. Dialogues are tracked by tagging utterances with addressee and mode information.

Speaker	addr	mode	text
Sgt	lt	normal	sir we should have first squad reinforce the lz
Medevac	lt	radio	eagle two six this is medevac two one
Medevac	lt	radio	turning final now
Medevac	lt	radio	have lz in sight
Lt	sgt	normal	roger. have first squad reinforce the lz

Figure 1: Example of MRE interaction

By collecting dialogues of a human trainee interacting with the virtual characters we can both evaluate and improve aspects of the system, such as speech recognition, natural language understanding, and dialogue interaction. For system evaluation, the MRE corpus is currently being coded for several phenomena at both utterance and subdialogue levels, tracking the lieutenant's attempted tasks, and response quality of the agents. (see (Traum et al., 2004) for more details). We also plan to study how human users react to virtual agents. We approach the analysis both quantitatively and qualitatively by using the coding schemes mentioned in section 4..

2.2. Military Radio

In order to investigate features of dialogue in this domain, we are building a corpus of human-human radio communication from military training exercises. The current state of the corpus is shown in table 2.

Name	Speakers	Utterances	Words
MZ	17	151	1912
SIM	38	2677	29,227
Total	55	2828	31,139

Table 2: Radio Data

The first activity (MZ) consists of short disjoint radio communication episodes from training exercises at Fort Leonard Wood. A much larger data set (SIM) consists of a single hour and twenty minute simulation exercise from Fort Rucker, involving trainees using flight simulators, exercising a coordinated mission with a command post and semi-automated simulated forces. This exercise contains communication on multiple radio channels involving often simultaneous speech, yielding nearly 10 hours total of speech, when separating out each channel. The recordings include both actual radio communication between physically separate people and communication between team members in the same location, captured over an open channel. A short fragment of the interaction is shown in Figure 2.

As can be seen in this fragment, not all sides of a dialogue occur on the same channel. The two members of the rogue 07 helicopter can be heard on channel 8. Channel 8 also captures their communication with other units, but this is also captured on other frequencies (45 for Savoy 06, and 42 for Rogue 06). The communication from Rogue 06 only appears on channel 42, so some of the content on channel 8 has only one side of a dialogue episode (which other bits are repeated on multiple frequencies). This situation presented a number of problems, including the need to correlate the data between eight channels, both to avoid redundancy of data, as well as to form a coherent picture of the multiple dialogues occurring at the same time. To track the dialogues, we code the addressee as well as mode (radio or normal), as in the MRE data. In the SIM data, there are also more complex issues of role and identity. Rogue Zero Seven (R07 in the example) is a team entity made up of two individual speakers, who speak to each other off radio but are generally indistinguishable to other radio callers (e.g. either may answer a call addressed to R07 or send a call as R07). Thus the corpus can be searched by individual speakers' dialogue, specific roles, or extended teams.

Though the addressee and mode coding allows most dialogues to be distinguished, the high number of speakers on some channels creates overlaps and affects dialogues that are otherwise distinct. For example in Figure 3, two people are calling ops at the same time for the same speaker, but are not in the same dialogue (and cannot hear each other when they are both speaking).

utt#	Spkr	text
1	DT	rogue zero seven,
2		[dragontoc ,]1
3		[uh / could we have the]2 ti:me for the: suspected border crossing .
4		over ,
5	P02	[dra:gono:ps ,]1
6		[this is predator zero two and zero one ?]2
7	R07	and ah: dragontoc ,
8		this is rogue uh zero seven .
9		uh say again ?

Figure 3: Example of SIM multiple dialogues

Thus dialogue episodes are tracked and episode information is coded on an utterance level. Episodes are also categorized by activity type classified according to the following 7 supercategories, which are based on the dialogue's general orientation to the task actions. Subcategories describe variations in dialogue formality, topic orientation and/ or speaker and addressee's orientation to the information (e.g. whether speaking of one's own action status or observed data). While some categories are specific to this

freq	Speaker	addr	mode	text
08	R07-A	R07-B	normal	okay . try uh: try to get savoy in ,
08, 45	R07-B	S06	radio	eh rogue+ correction ah , savoy zero six . this is rogue / zero seven ? //
08	R07-A	R07-B	normal	wait where did they go: . lost them . //
42	R06	R07	radio	seven , this is six . did you notice that they (reset) us ?
08, 42	R07-A	R06	radio	roger , ah / you back at the / the / station ? or what .

Figure 2: Radio data sample dialogue

domain (e.g. 'Radio Check'), most of the domain-specific information occurs in subcategories, while the supercategories are transferable across domains.

1. Radio Check
2. Task Allocation: Orders, Action Prompt, Negotiating Task
3. Status Report: Call, Action Status, Narrating
4. Information Sharing: Spot Report, FYI, Advising
5. Information Gathering: Request Action Status, Data Request, Procedural Clarification Request
6. Achieving Task: Action Clearance, Coordinating Action, Problem Solving
7. Socializing: Storytelling, Commenting

3. Preliminary Results

Table 3 shows the amount of transcribed data at this stage of the development of the corpus, as well as the size of the total vocabulary for MRE agents vs. humans and SIM data.

Type	Speakers	Utterances	Words	Unique
MRE Agents	8	2,373	11,766	222
MRE Human	22	1,480	6,833	466
SIM	38	2,677	29,227	1,567

Table 3: MRE and SIM by utterances, words, and unique words

In MRE, the Agents spoke 11,766 words, although their total used vocabulary size was only 222 distinct words. Though we have had 30 tests, the human vocabulary is only twice as big, reflecting the narrow focus of the scenario. Although we may expect some vocabulary increase with subsequent tests, increase in vocabulary has declined to almost zero in our most recent test runs, as shown in figure 4.

The human data in the MRE vs Radio corpus displays both similarities and differences due to the features of the domains and modalities. Utterances in the MRE tests tend to be much shorter, with an average length of 4.6 words per utterance as opposed to 11 in the radio data. This coincides with the different intonation data in table 4

The breakdown of intonation types show that falling and continuing intonation is predominant overall, which means that we may expect many questions being uttered with non-rising intonation. In the MRE data the falling intonation

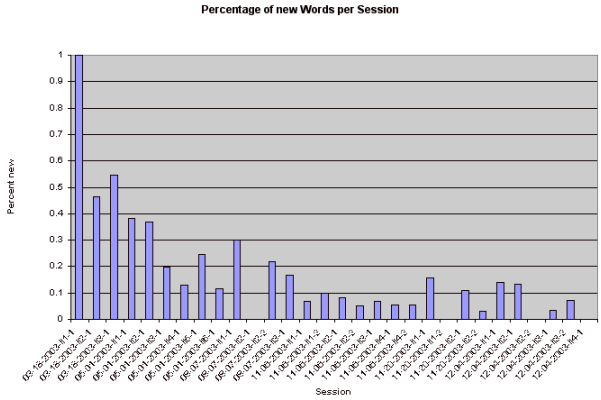

Figure 4: Percentage of new words per MRE session

Intonation	Radio	MRE
Rising	904 (11.63)	113 (5.91)
Continuing	3176 (40.85)	406 (21.25)
Falling	3694 (47.52)	1392 (72.84)
Total	7774	1911

Table 4: Radio and MRE data by Intonation Type

dominates much more than in the radio data, in part due to the shorter utterance lengths, also possibly a higher number of command statements. A co-occurrence study with the types and amount of core acts (see section 4.) will help us build more theories on the character of the data.

The lexical patterns of each corpus are suggestive of broader trends in the data as well. In the MRE domain, which is a relatively brief and enclosed scenario, words for entities (sergeant, lz, squad, tucci, boy) and actions (secure, send) are in the top ten rank frequency. In the longer, more complex SIM domain, elements of call signs (zero, six, two, rogue, roger) are among the most frequent items, as they play an important role in radio communication (Martinovski et al., 2003).

Table 5 compares the frequency of the top ten and selected words in spoken English (from British National Corpus http://www.comp.lancs.ac.uk/ucrel/bncfreq/flists.html) with the SIM training simulation data and the human utterances from MRE.

There are several striking patterns of frequency differences in common words that seem to support some of our intuitive characterization of the domains.

Team oriented data: Though 'we' is slightly more common in our data, 'I' is significantly less common in both corpora, 10 times less frequent in MRE than BNC.

Word	BNC		SIM		MRE	
	#	freq	#	freq	#	freq
the	1	3.961	5	2.922	1	10.22
I	2	2.945	23	1.187	70	0.278
is	3	2.784	2	3.425	3	3.908
you	4	2.596	11	2.108	1	21.04
and	5	2.521	17	1.512	30	0.746
it	6	2.451	31	0.674	77	0.263
to	7	2.186	6	2.522	5	3.322
that	8	2.150	20	1.351	26	0.834
a	9	1.864	22	1.204	17	1.215
er[1]	10	1.457	4	3.100	89	0.190
of	11	1.455	27	0.780	24	0.849
we	14	1.045	10	2.203	14	1.551
they	16	0.933	37	0.554	128	0.088
are	17	0.892	15	1.783	45	0.659
was	18	0.810	72	0.240	128	0.088
what	21	0.731	53	0.356	7	2.883
he	22	0.728	-	0.079	89	0.190
this	27	0.563	9	2.355	21	1.039
will	32	0.496	40	0.479	104	0.146
where	-	0.154	-	0.200	28	0.790

Table 5: Rank and frequency per 100 words in British National Corpus Spoken Data, SIM and MRE human data

Task/Simulation Immersion: 'is' roughly equal to BNC, but past lexeme 'was' is several times less frequent in SIM and nearly 10 times less frequent in MRE. An indicator of future planing 'will' (also includes 'll) is also strongly less frequent in MRE, perhaps due to the limited time frame and immersion in events directly surrounding the Lt. SIM, which has much more complex interactions between teams coordinating to fulfil their missions has a comparable level.

Pronouns in general (excepting 'we') are less frequent in our data, particularly 3rd person reference, perhaps again due to the immersion in the present task at hand, but also likely due, especially in the MRE dialogue, to short statements. The lower frequency of 'and' in the MRE data and the much higher occurrences (over 10%) of 'the' support this observation as well. The MRE utterances also appear more controlled and carefully spoken, given the extremely low occurrence of hesitation sounds. Finally, 'what' and 'where' occur with much higher frequency in MRE than in the other data. It is unclear at this point whether this is a result of the specific scenario (where the lieutenant meets his platoon mid action) or possibly a more general feature of simulated experience, where the human participant relies on the agents for information to help him immerse more fully in their world.

4. Future Plans

The result of the domain specific annotations described above is a corpus that may be searched by utterance, intonation unit, speaker or entity, or by subdialogue activity. Currently we have coded a portion of SIM and will expand to MRE utterance unit level coding for more productive co-occurrence studies. These coding categories include grammatical coding (sentence structure and ellipsis), reference coding, a wide range of dialogue acts (e.g. suggestion, request, statement, answer, politeness), grounding acts (Traum and Allen, 1994), and communication management (interactive reformulation, hesitation sound) (Martinovski, 2001)

[1]Includes "erm" "uh" and "um".

Acknowledgements

We would like to thank many members of the MRE project team for help in this work. Additionally we would like to thank John Lowry and David Dunstedter for making available the Radio training data. The work described in this paper was supported by the Department of the Army under contract number DAAD 19-99-D-0046. Any opinions, findings and conclusions or recommendations expressed in this paper are those of the authors and do not necessarily reflect the views of the Department of the Army.

5. References

Carletta, J., A. Isard, S. Isard, J. Kowtko, G. Doherty-Sneddon, and Anne Anderson, 1996. HCRC dialogue structure coding manual. Technical Report 82, HCRC.

Heeman, Peter A. and James Allen, 1994. The TRAINS 93 dialogues. TRAINS Technical Note 94-2, Department of Computer Science, University of Rochester.

Jurafsky, Daniel, Rebecca Bates, Noah Coccaro, Rachel Martin, Marie Meteer, Klaus Ries, Elizabeth Shriberg, Andreas Stolcke, Paul Taylor, and Carol Van Ess-Dykema., 1998. Switchboard discourse language modeling project final report. Research Note 30, Center for Speech and Language Processing, Johns Hopkins University.

Martinovski, Bilyana, 2001. *The Role of Repetitions and Reformulations in Court Proceedings – a Comparison of Sweden and Bulgaria*. Ph.D. thesis, Göteborg University: Department of Linguistics.

Martinovski, Bilyana, David Traum, Susan Robinson, and Saurabh Garg, 2003. Functions and patterns of speaker and addressee identifications in distributed complex organizational tasks over radio. In *Diabruck: seventh workshop on semantics and pragmatics of dialogue*.

Swartout, William, Jonathan Gratch, Randall W. Hill Jr., Eduard Hovy, Richard Lindheim, Stacy Marsella, Jeff Rickel, and David Traum, 2004. Simulation meets hollywood: Integrating graphics, sound, story and character for immersive simulation. In Oliviero Stock and Massimo Zancanaro (eds.), *Multimodal Intelligent Information Presentation*. Kluwer.

Traum, David, Susan Robinson, and Jens Stephan, 2004. Evaluation of multi-party virtual reality dialogue interaction. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC)*.

Traum, David R. and James F. Allen, 1994. Discourse obligations in dialogue processing. In *Proceedings of the 32nd Annual Meeting of the Association for Computational Linguistics*.